Chapter Two
Administrative Data and Record Linkage Issues

For research purposes, administrative data have the advantage of detailed and accurate measurement of program status and outcomes, complete coverage of populations of interest (enabling detailed subgroup analyses), data on the same individuals over long periods, low cost relative to survey data, and the ability to obtain many kinds of information through matching (Hotz et al., 1998). In addition, many types of administrative data have relatively high degrees of uniformity in content across geographic areas.[13,14]

Government agencies have recognized the potential research uses of administrative data. Of 10 data development initiatives recently identified by USDA's Economic Research Service, only one did not involve administrative data (Wittenburg et al., 2001). Five of the 10 initiatives involve creation of linked databases matching administrative records from multiple agencies, or matching administrative records to survey responses.

Linked databases are a way to create "new" data from existing sources. For example, the National Center for Health Statistics (NCHS) determines infant mortality rates using state files of linked data from birth and death certificates (Mathews et al., 2002). The Department of Transportation examines motor vehicle crash outcomes by linking records of police-reported crashes to hospital discharge data, EMS data, and hospital emergency department data (NHTSA, 1996a). In the social services arena, a number of States have developed master client indexes that match administrative records from multiple social service programs to obtain unduplicated counts of clients and examine patterns of multiple program use (UC Data, 1999).

In a recent report, the U.S. General Accounting Office (GAO, 2001) noted that: "Federally sponsored linkage projects conducted for research and statistical purposes have many potential benefits, such as informing policy debates, tracking program outcomes, helping local government or business planning, or contributing knowledge that, in some cases, might benefit millions of people." The GAO also noted that record linkage projects generally raise significant concerns about privacy protection because "person-specific data are involved and because actual linkages typically occur at the individual level, multiplying the quantity of data recorded on each individual." But the GAO concluded that various techniques may help address privacy concerns (such as signed consent forms, masked data sharing, and secure data centers) and strategies for enhancing data stewardship could help ensure the confidentiality and security of linked data.

This chapter discusses research uses of administrative data, methods of implementing record linkage, and issues that must be considered in planning or implementing record linkage systems.

---

[13] Many data elements in State administrative systems are required to meet federal regulations. The result is content uniformity, even though the data systems may vary in structure and format.

[14] Administrative data have some disadvantages: the data can be costly to collect and process; for some purposes, administrative data may be missing many data elements of interest and some data elements may have considerable measurement error; and administrative data are not easily accessed by researchers.

Administrative Data

Administrative data are the data assembled for program operations. Data for individual program participants are maintained in management information systems designed to determine eligibility and benefits at application, collect participant characteristics for reporting purposes, maintain histories of benefit receipt, and, in the case of WIC, track client activities such as referrals and appointments for nutrition education.

Administrative data systems for social service programs have become more complex over time. In the past decade, two pieces of federal legislation put increased demands on data systems. The *Personal Responsibility and Work Opportunity Act of 1996* (PRWORA) replaced Aid to Families with Dependent Children (AFDC) with TANF and introduced work requirements and time limits for some participants in the FSP. Both TANF and FSP information systems now track longitudinal data in order to implement rules and monitor compliance. The *Government Performance and Results Act of 1993* (GPRA) requires government agencies to develop strategic plans with measurable goals. GPRA requirements place demands on administrative data systems to monitor progress against performance goals. For example, USDA's Food and Nutrition Service strategic plan includes a goal to increase breastfeeding initiation among WIC participants (USDA/FNS, 2000), and WIC administrative data systems were modified in 1998 and 2000 to provide data on breastfeeding initiation that is consistent across time and across State agencies.

Research Uses of Administrative Data

Historically, administrative data from the FANPs have been used for a variety of research purposes. Administrative data are used to periodically take stock of the number and characteristics of program participants. For example, the biennial *WIC Participant and Program Characteristics Studies* (PCs) (Bartlett, et al. 2002) are based on administrative data collected from State WIC agencies, and the annual *Characteristics of Food Stamp Households* (Tuttle, 2002) are based on FSP administrative data assembled for quality control purposes.

Administrative data are also regularly used for program evaluation. USDA evaluation studies have used administrative data to create sample frames for surveys and to examine a wide array of program operation and program outcome issues. These studies, however, are one-time evaluations and the scope of data collection and analysis is sometimes limited to a single application.

Research uses of FANP administrative data are paralleled by other social service programs. The University of California (UC Data, 1999) conducted an inventory of research uses of administrative data and found over 100 examples of research uses of administrative data among social service programs. The 100 examples were found across the substantive areas of welfare experiments, child welfare research, and health care research. Many of these examples were one-time evaluations.

The UC Data report highlighted efforts to link databases for research and evaluation, or to create ongoing data systems to enhance the reporting capabilities of administrative data. Three linkage strategies were identified:

- Data integration—multiple data systems are integrated on the same computer hardware, or through data exchange in real-time.
- Computer matching—personal identifier (usually SSN) is used to retrieve data from external databases through batch merges or ad-hoc queries.

- Record linkage—data extracts from multiple systems are combined to create a new database (data warehouse).

Data integration and computer matching are techniques applied to internal operations and usually arise to support program operations (data integration streamlines operations and computer matching enables data verification). The end result is an administrative database with enhanced capability to meet research needs. Record linkage, on the other hand, generally occurs outside of normal program operations by a research division or external research entity for the primary purpose of enhanced reporting and research capabilities.[15] Examples of each of these techniques are discussed below.

Data Integration

The most common example of data integration cited by UC Data is the integration of AFDC/TANF, food stamps, and Medicaid data systems. In 1998, 20 of the 26 States surveyed by UC Data had integrated systems for these three programs; 12 of the integrated systems also included the Job Opportunities and Basic Skills (JOBS) program. Programs that were less commonly integrated into these systems were: Child care subsidies (3 States); Foster Care (2 States); Child Support Enforcement (1 State); General Assistance (1 State); and Child Protective Services (1 State).

Data integration enables direct measurement of multiple program participation from a single client database. For example, in integrated systems, food stamp cases are automatically denoted FS-PA (food stamps and public assistance) or FS-NPA (food stamps and no public assistance) according to the case status in public assistance programs (TANF, SSI, and general assistance). Longitudinal case histories from the single data system can be examined to determine whether the dynamics of FSP and TANF entry and exit coincide.

Computer Matching in the Food Stamp Program

The FSP uses computer matching to improve program efficiency and integrity. Federal regulations require FSP applicants to provide their Social Security number (SSN) (7 CFR 273.6) and regulations authorize State FSP agencies to use SSNs to routinely match FSP participant records to external data systems.

State food stamp agencies perform computer matches for three main purposes: to identify ineligible participants, detect dual participation, and verify eligibility. Ineligible participants are identified by computer matches with the Social Security Administration (SSA) Death Match file and the Prisoner Verification System are done to identify ineligible participants. Dual participation is detected through computer matches with FSP data systems in neighboring States. And eligibility is verified through computer matches to external databases to verify information provided by participants during the certification process (State Wage Information Collection Agency (SWICA), State Data Exchange (SDX), Unemployment Insurance (UI), and Beneficiary Data Exchange (BENDEX)).[16] Currently, the only computer matches that are mandated for FSP agencies are matches to the SSA Death Match file and Prisoner Verification System (USDA/FNS, 2002).

---

[15] Databases created through record linkage have limited potential to serve operational needs because the databases are generally not updated in real-time.

[16] These data systems are all part of the Income Eligibility and Verification System (IEVS). IEVS was mandated for use by the FSP, prior to 1996. PRWORA (1996) removed the mandate but IEVS continues to be used because these systems are perceived to provide useful data (USDA/FNS, 2002).

USDA found that use of computer matching by State FSP agencies almost doubled in the decade from 1991 to 2001—from an average of 7.5 matching systems used per State, to 14 (USDA/FNS, 2002). In addition, increases in computer processing capacity and growth in communications networks led to a transition from batch processing to real-time links between FSP data systems and external databases.

Computer matching typically involves transmission of data from one agency to another, with a "result code" returned to indicate the quality of the match. Computer matching, as used by FSP agencies, does not pull source data from an external database to add to the primary database. Use of computer matching for program operations demonstrates the technological feasibility of linking large separate data systems by use of a single, unique, verified identifier (SSN).

Record Linkage Projects within the Social Services

Record linkage projects join records from two or more separate data systems to create a new record in a new database. [17] Two recent studies provide numerous examples of record linkage projects. The UC Data inventory of administrative data systems cites examples of record linkage from welfare demonstration evaluations and from State projects creating "master client indexes" of social service clients. GAO (2001) provides examples of record linkage projects conducted under federal auspices or with federal funding.

Many welfare evaluation studies created linked databases to join information about program participation to outcomes data on employment and earnings. For example, the Alabama ASSETS demonstration project in the mid-1990s linked monthly AFDC, Food Stamps, JOBS, child support, and UI earnings data to create linked longitudinal databases. Similarly, the Florida Family Transition Program (FTP) demonstration study linked data extracts from AFDC/TANF/FSP to Department of Labor quarterly earnings records, Medicaid claims, and childcare subsidy records. However, linked databases from welfare evaluations were created at a point in time and do not support ongoing reporting.

Much interest has been generated in recent years from development of data warehouses that link data from multiple social service programs on an ongoing basis. Linked databases have been developed under the auspices of State Departments of Health to provide improved data access and data quality to State agencies responsible for surveillance, research, and program planning. In 1999, UC Data found that five States were developing or operating state-level master indexes of social service clients.[18] Linked databases appear in some cases (Texas is discussed below) to provide an interim solution on the way toward fully integrated data systems for all social service programs.

Record linkage across many social service programs is more difficult to achieve than computer matching based on verified SSN. Many programs do not collect or do not verify SSNs and, as a result, record linkage must rely on personal identifiers (name, date of birth, gender, race, address, phone) that are not unique and are subject to change over time. While the UC Data study cites several

[17] Pioneering work on record linkage was done by Newcombe in the 1950s in the area of health research (Newcombe et al., 1959).

[18] Texas was operating The Integrated Database Network (IDBN); Washington was operating the Client Services Database (CSD); South Carolina maintained a data warehouse called the "master file" that brought together data from the separate FSP, TANF, and Medicaid systems in that State; and Tennessee and Minnesota had data warehouse projects under development.

examples of record linkage projects, for the most part, that study did not indicate the methods used to link data. Methods may be as simple as a merge on shared program ID or SSN, or as complex as probabilistic record linkage (these methods are discussed below).

One of the first efforts at an integrated cross-agency database constructed by probabilistic record linkage is the Illinois Integrated Database (IDB) on Children's Services developed by the Chapin-Hall Center for Children at the University of Chicago (Goerge et al., 1994; Goerge, 1997). Development of this database began in the early 1980s with construction of a longitudinal foster care database to study foster care dynamics. Currently, the database contains data from eight social service agencies and documents all contacts that a child has with TANF, Medicaid, food stamps, child welfare, special education, corrections and juvenile justice, mental health, and substance abuse (Goerge and Van Voorhis, 2002).

The Chapin-Hall database exists outside the Illinois State agency information system, and maintains longitudinal case histories. In 1997, the Illinois Department of Human Services (DHS) implemented the Common Client Index containing an unduplicated list of recipients of all DHS services; this system contains the most recent information about a client but does not contain case histories (UC Data, 1999).

Other States have developed master client indexes that have evolved over time. For example, UC Data reported on the Washington State Department of Social and Health Services (DSHS) development of the Needs Assessment Database. This database was developed in 1990 to determine the number of clients served by multiple agencies within DSHS. The database combined data extracts from 15 agencies to determine the number of shared clients and the total costs accrued for shared clients. The effort was a point-in-time linkage of cross-sectional data extracts, and was repeated in 1992 and 1994. In 1996 this database evolved into the Client Services Database (CSDB) which links extracts on a more frequent basis.

In Texas, the Integrated Database Network (IDBN) was implemented in 1995, linking data from four agencies with separate data systems. UC Data reported that the IDBN was developed for two distinct purposes: to assist workers in the field to rapidly collect information on clients necessary for case processing, and to assist state agency staff in statistical and management reporting. The system was designed to eventually link data from all eleven agencies within the Department of Health and Human Services. IDBN, however, will be superseded by the Texas Integrated Eligibility Redesign System (TIERS) project, launched by State legislation in 1999. TIERS will be developed as a fully integrated eligibility and enrollment system to include services provided by the Texas Department of Human Services (Food Stamps, TANF, Medicaid, Children’s Health Insurance Program (CHIP), Refugee Assistance, Community Care for the Aged and Disabled, and Hospice) and support for sharing data with other State agencies (TDHS, 1999).

Record Linkage Methods and Issues

Record linkage and computer matching are terms that refer to a process of matching records from different data files — from multiple data systems or from the same data system at different points in time. Computer matching typically refers to the process of matching (or verifying) specific information with an external file and adding a result code to the primary file indicating the quality of the match. Record linkage typically describes a process that links records from more than one file and returns a new record for a completely new data file.

Types of Record-Linkage

There are three methods of record linkage: match-merge, deterministic linking, and probabilistic linking (Whalen et al., 2001). A match-merge relies on an exact match of a single common identifier present in two files. Deterministic record linkage requires an exact match of identifying information, but uses multiple criteria to establish a match. Probabilistic record linkage is made when the calculated statistical probability of a match exceeds a certain threshold.

Match-merge techniques are generally used only when information originates from the same data system or when identifiers (such as SSN) are very reliable.[19] For example, a match-merge may be used to link FSP participants in data extracts drawn at different points in time, with participants linked by the FSP system ID. A match-merge will fail in this case only for participants who exit and re-enter the system with new IDs.

Deterministic record linkage uses multiple criteria to establish a match between records. For example, the link might require a match on SSN *or* name and date of birth. Multiple criteria introduce the complication that data items vary in quality or reliability. Match routines use information about the varying quality of data items, either explicitly or implicitly. Some applications sequentially test multiple deterministic criteria, excluding matches at each step from the next step of matching. Information about quality of data items is used to establish the ordering of criteria. Alternatively, several criteria could be applied at the same time, with points assigned to each criterion and a point threshold used to establish a match. Assigning different points to different identifiers provides a way to recognize variations in quality or reliability of different data items.[20]

Probabilistic record linkage identifies a match between records based on a formal statistical model. The advantage of probabilistic record linkage is that it uses all available identifiers to establish a match (e.g., name, sex, date of birth, SSN, race, address, phone number) and does not require identifiers to match exactly. Identifiers that do not match exactly are assigned a "distance" measure to express the degree of difference between files. Each identifier is assigned a weight and the total weighted comparison yields a score, which is used to classify records as linked, not linked, or uncertainly linked according to whether the statistical probability of a match exceeds a certain threshold (Winkler, 1999).

Probabilistic record linkage models were first introduced by Newcombe (1959) and formalized by Fellegi and Sunter (1969). Modern probabilistic record linkage is a collection of techniques from computer science, statistics, and operations research (Winkler, 1994). These techniques include string comparison methods, algorithms for scaling commonly occurring values, and methods for scoring the comparisons of multiple identifiers and assigning a match probability to the total score. Probabilistic methods provide the most accurate means of matching files that do not share a single common identifier.[21]

---

[19] Reliability of the single identifier must be comparable across the files being matched. For example, a match merge on SSN across the FSP, which verifies SSN, and another program, which does not verify SSN, may result in large numbers of false positive and false negative matches.

[20] An example given by Whalen (2001) requires a total point score of 25 or greater to establish a match, with points assigned as follows: 20 points for SSN agreement, 15 points for last name agreement, 8 points for first name agreement, 5 points for date of birth agreement, 1 point for gender agreement and –10 points if gender does not agree.

[21] One validity study compared Statistics Canada's linked birth and infant death records to hospital records and found "a high degree of agreement … suggest(ing) a high degree of validity" (Fair, et. al, 2000).

Record-Linkage Issues

Deterministic and probabilistic record linkage methods are used to link databases that lack a unique and reliable common identifier. If SSNs are present on the databases of all social service programs, and are verified at application, record linkage could be achieved by a simple match-merge. In reality, however, SSNs are not used by all social service agencies, and SSNs are not always verified when they are collected.

The success of deterministic and probabilistic record linkage depends on common identifiers, standardized data fields, and data retention that ensures that contemporaneous data are available for the files being linked. Identifiers are data items that identify an individual — first and last name, SSN, date of birth, race, gender, address, phone. Common identifiers must be present in the files to be matched and they must appear in the same format.

Data standardization involves recoding categorical data items and standardizing the structure and content of data fields. Categorical data items, such as race and gender, will not match across files if based on different coding schemes (e.g., GENDER may be coded as 1/2 or M/F for male/female). Imposing a consistent coding scheme is usually a simple matter of recoding variables in some of the files being matched.

Standardizing data fields that are not categorical, such as name and address, often requires parsing data items and translating the contents of data fields. For example, if a NAME field contains first and last name, it must be parsed to separate fields (FNAME, LNAME) to enable separate matching of first and last name. It may be desirable to translate the content of name fields to increase the likelihood of matches; for example, by replacing all nicknames with formal names or removing all titles (Mr., Mrs., Jr.). With address fields, content translation is imperative to eliminate variations that would preclude a match. Typically all spelling variations on street types (Avenue, Boulevard, Circle, Highway, Road, Route) and prefix/suffix direction on street names (East, West) are translated to standard Census abbreviations prior to matching. Address data must also be parsed into separate fields (house number, street name, street type, directional prefix/suffix) to enable separate comparisons of comparable data fields.

Data retention refers to retention of information when individual data fields are updated to reflect change. Most personal identifiers are subject to change over time — names change due to marriage, divorce, or adoption; addresses and phone numbers change due to relocation; ZIP Codes may change due to reassignment by the postal system. Two data files extracted from separate data systems at the same point in time may contain information on the same individual *entered* at different points in time. Probabilistic record matching can incorporate "old" information by testing for a match on every combination of current information and old information across two data files.

Methods of Implementing Probabilistic Record Linkage

Probabilistic record linkage has been implemented in record linkage software systems that are available commercially and from government agencies (Winkler, 2001). Current record linkage systems are described below with examples of their application.

The Department of Transportation's Crash Outcome Data Evaluation System (CODES) links records of police-reported motor-vehicle crashes to hospital discharge data, Emergency Medical System (EMS) data, and hospital emergency department data. The system was developed in response to a

Congressional mandate to determine the benefits of safety belt use and motorcycle helmet use.[22] The CODES system was implemented in seven States in 1996 and the National Highway Traffic Safety Administration (NHTSA) has since funded the system in an additional 20 States.[23] The system uses commercial AutoMatch software, which is no longer available under the AutoMatch name. AutoMatch was acquired by Vality Technology, which is now a part of Ascential Software; this matching software has evolved into part of the Integrity enterprise solution product.[24]

The Master Child Index (MCI) being developed by the City of New York, Department of Health links records from the Citywide Immunization Registry (CIR) and Lead Poisoning Prevention Program (LPPP) to facilitate the identification and tracking of children for immunizations and lead screening. In April 2002, the ChoiceMaker commercial software was chosen to implement record linkage.[25] ChoiceMaker Technologies® was established in 1998 and has developed matching software with partial funding from the National Science Foundation.

The Integrated Data Base developed by the Substance Abuse and Mental Health Services Administration (SAMHSA) of the U.S. Department of Health and Human Services used probabilistic record linkage to link client records from three agencies in each of six States. The agencies were Medicaid, State mental health, and State substance abuse agencies. The integrated database was built with 1996 data and supported research on treatment services received from each type of agency (Coffey, et al., 2001). Record linkage was implemented by a system of SAS® programs; these programs are available on the SAMHSA web site. [26]

Statistics Canada and the U.S. Bureau of the Census use record linkage for population enumeration operations. The software used by Statistics Canada is CANLINK; this software contains record linkage operations but does not perform name or address standardization (Winkler, 2001). The U.S. Bureau of the Census uses software for name standardization, address standardization, and record linkage. The Census software was written in C++ and the compiled code runs on all computers. Source code and documentation for the Census programs are available, but not supported (Winkler, 2001).

While record linkage software is available, Winkler cautions that "record linkage is like messy data analysis … individuals need to recognize patterns in data" and "groups undertaking matching must be aware of the large amounts of time and resources needed for developing person skills and for cleaning up lists" (Winkler, 2001). Phase II of this project will investigate the SAS system developed for SAMHSA and the Census software, for application to FANP data.

---

22 Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991.

23 Information is available at www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/CODES.html.

24 A discussion of the original AutoMatch software can be found in Jaro (1995). Information about Integrity is available at www.vality.com. Winkler (2001) cites the price of Integrity as $195,000 plus 15% maintenance.

25 Information about ChoiceMaker is available at www.choicemaker.com.

26 The system contains 6 primary SAS programs and 23 SAS macros. The programs are available at www.samhsa.gov.